Assigning Geographical Scopes To Web Pages*

Bruno Martins, Marcirio Chaves and Mário J. Silva

Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa
1749-016 Lisboa, Portugal

1 Introduction

Finding automatic ways of attaching geographical scopes to on-line resources, also called "geo-referencing" documents, is a challenging problem, getting increasing attention [1, 5, 3]. Here we present a system architecture and a process for identifying the geographical scope of Web pages, defining a scope as the region where more people than average would find that page relevant. We rely on typical Web IR heuristics (i.e. feature weighting, hypertext topic locality, anchor description) and assumptions on how people use geographical references in documents. The method involves three major steps. First, geographical named entities are identified in the text. Next, we propagate the found named entities through the Web linkage graph. Finally, a geographical ontology is used to disambiguate among the named entities associated to a document, this way selecting the most likely scope. In the future, we plan on using scopes in new location-aware search tools.

2 System Architecture

The proposed architecture relies on Semantic Web standards such as RDF and Dublin Core. Documents are harvested into XMLBase, our Web data management system which contains a crawler, data/meta-data repositories, and several document analysis components (i.e. language identification, document parsing). The sequential processing stages of the scope assignment algorithm take RDF representations of the documents and augment them with additional information. In the end, a geographical scope is assigned to each document. The geographical information used by the algorithm is kept in GKB [2], a common knowledge base integrating data from multiple external resources (i.e. public gazetters and databases). GKB essentially includes place names and the ontological relationships between them (i.e. broader/narrower geographical entities), supporting mechanisms for storing, maintaining and exporting this information.

3 Step 1 - Geographic Named Entity Recognition

After low level document processing operations (text extraction and tokenization) we identify the geographical named entities (toponyms) present in the text through a simple named entity recognition (NER) approach. This is based on list lookups (using place

* This research was partially supported Fundação para a Ciência e Tecnologia, under grants POSI/SRI/40193/2001 and SFRH/BD/10757/2002.

names from GKB) and heuristics such as capitalization and surrounding text. Surrounding text is also used to disambiguate the types of places mentioned in the text (i.e. "city of Lisbon", "Setubal district"). Mikheev et al. showed that a NER system could perform well without gazetteers for most entity classes, but not for place names [6]. The same study found that simple list lookup performs reasonably well for locations. Previous studies have also shown that geographic name types are commonly disambiguated in the text itself [4]. The discovered named entities are weighted according to their occurrence frequency and HTML markup information (i.e. text from the title of the documents or from hypertext anchors is considered more important). These weights are used afterwards as disambiguating properties, since the same document may reference several different geographic entities.

4 Step 2 - Web Graph Propagation

Weights for the entities recognized in document d are divided between all linking documents associated with d in the Web graph ($\frac{d}{|inlinks(d)|}$). The value is then assigned to the same entity in the linking documents. This "propagation" procedure is applied only once for each page, and therefore entities contribute only to pages that are one hyperlink away from the source. We also use heuristics to guide this propagation stage. For instance, documents hosted on the same site are considered more likely to relate to the same geographical concept, and therefore weights propagated through these links are given extra credit.

5 Step 3 - Assigning Scopes to Documents

Named entity recognition in itself does not derive the meaning of the expressions recognized. A major problem concerns ambiguity, as for instance "Odivelas" refers to both a city near "Lisbon" and another in "Alentejo" (the referent ambiguity problem). The same location can also have more than one name (reference ambiguity) and this latter problem has another twist: the same name can be used for locations as well as for other class of entities such as persons (referent class ambiguity). Our final stage involves disambiguating the entities associated with a document from the previous steps, and the semantic relationships between these entities, to decide (if possible and reasonable) on the scope to be assigned to each page. Information from the GKB is used to build a geographical ontology (essentially a place hierarchy). Each node is "activated" with the weight associated to its defining named entity in the document. These values are then propagated across the ontological relationships between the entities, using inference methods from probabilistic graphical models. For instance, if "Lisbon" is associated with a page, some weight is also given to all the entities corresponding to sub-regions of "Lisbon". Again, heuristics are used to guide the process, as different ontological relationships (i.e. narrower/broader, equivalent) propagate weights in different ways. Finally, we select the highest weighted entity as the most probable scope for the document, or none if all entities are weighted below a given threshold.

6 Conclusions

Statistics collected through our Web search engine tumba! (`www.tumba.pt`) motivated this research, in the sense that geographic information is pervasive on both documents and queries. A prototype system currently implements most of the ideas described here, and we are now starting evaluation experiments in tandem with additional software development. Since many parameters are combined, a very important step concerns tuning the "importance" given to each of them. For now, we are essentially relying on empirical tests and on published results from other IR experiments. In the future, we plan on separately evaluating the different aspects involved in our approach. The geographical named entity recognition step will be more thoroughly evaluated through the participation on a joint evaluation promoted by Linguateca (`www.linguateca.pt`). As for the evaluation of the system as a whole, we intend to use a test collection built from the following sources:

- Pages from sites for Portuguese municipalities, under the assumption that all the pages in a site belong to a geographic scope covering the area of the municipality.
- Pages from the Open Directory Project located under the branch devoted to Portuguese pages with a coherent geographic scope.
- Pages under the RCTS network (public infrastructure hosting sites for schools, museums, and other institutions) under the assumption that all the pages in the same Web site belong to the geographical scope of the institution. From these pages, we only considered the ones from public schools, where the scope is well defined.

Although this collection cannot accurately model the linkage information found on a large sample of the Web, it will nonetheless allows us to automatically test the algorithm on a relatively large sample of Web pages.

References

1. E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: Geotagging Web content. In *Proceedings of SIGIR-04, the 27th annual international conference on Research and develoepement in information retrieval*, pages 273–280. ACM Press, 2004.
2. M. Chaves, B. Martins, and M. J. Silva. Grease Knowledge Base. DI/FCUL TR 04–XX, Department of Informatics, University of Lisbon, November 2004.
3. J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *Proceedings of VLDB-00, the 26th International Conference on Very Large Data Bases*, pages 545–556. Morgan Kaufmann Publishers Inc., 2000.
4. L. L. Hill, J. Frew, and Q. Zheng. Geographic names - the implementation of a gazetteer in a georeferenced digital library. *D-Lib Magazine*, 5(1), January 1999.
5. C. B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. van Kreveld, and R. Weibel. Spatial information retrieval and geographical ontologies: An overview of the SPIRIT project. In *Proceedings of SIGIR-02, the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 387–388. ACM Press, August 2002.
6. A. Mikheev, M. Moens, and C. Grover. Named entity recognition without gazetteers. In *Proceedings of EACL-99, the 9th Conference of the European Chapter of the Association for Computational Linguistics*, 1999.